Technology in documentation

TEI and the Nxaʔamxcín Dictionary

Ewa Czaykowska-Higgins and Martin Holmes
University of Victoria

Acknowledgements

- The late M. Dale Kinkade
- The speakers who worked with Dale in the 1960s & 70s
- Mary Marchand, Elizabeth Davis, the late Agatha Bart and the late Tillie George who urged us to work on a dictionary in 1990
- Mary Marchand and Pauline Stensgar who continue to support this project
- The various students who worked on inputting the materials into Lexware and XML, and programmers Bob Hsu and Greg Newton
- Colville Tribes Culture Committee
- Social Sciences and Humanities Research Council of Canada, University of Victoria Faculty of Humanities, and Humanities Computing and Media Centre

Outline of Presentation

Part I Background: The Project and The Nxaʔamxcín Language

Early Beginnings of the Project: From Filecards to Lexware

Part II Using TEI for a Lexical Resource Project

Evolution and Early Challenges

From Lexware to Text-Encoding Initiative structure

Describing what TEI can do

Part III Comparing Standards

Format and Terminological Interoperability

Part I

Background The Project and The Nxaʔamxcín Language

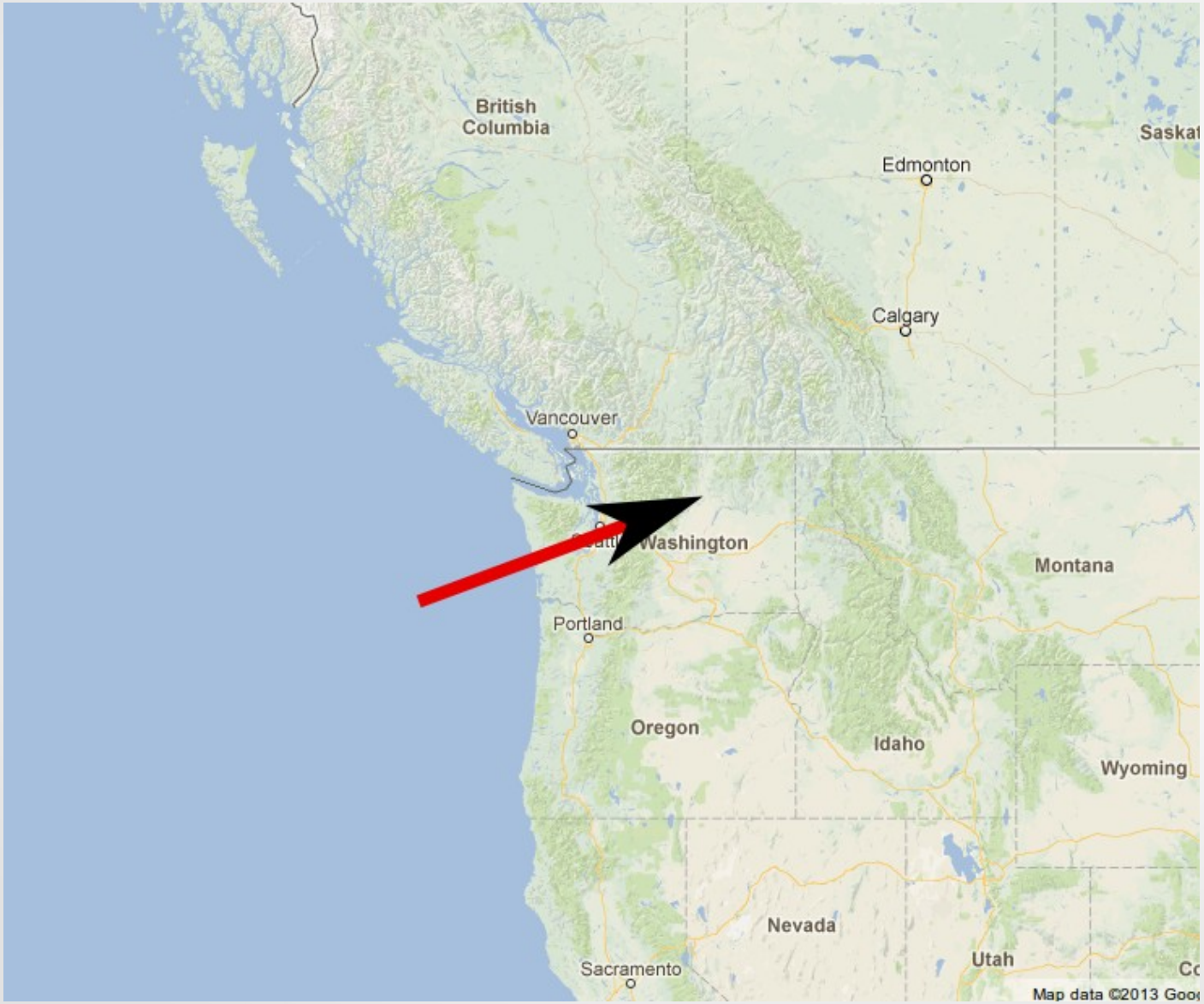
The Project Today

- To construct and make publicly available
  - Rich and extensive lexical documentation of Nxaʔamxcín
  - Recorded in the 1960s and 70s with approx 22 speakers, some in their 80s and 90s at the time
  - Therefore: legacy materials
- To create from these legacy materials lexical resources
  - That are both enduring and reusable in various formats and for various purposes
  - Including as a print dictionary and as a searchable, extendable web-based database which can be used in different views, for learning, teaching, language revitalization, and scholarly purposes

The Project Team

- At Colville Tribes
  - Elders: Pauline Stensgar, Mary Marchand
  - Language Program: Ernest K'saw's Brooks, Sharon Covington; Albert Andrews
  - History and Archeology: Guy Moura
- At UVIC
  - Martin Holmes, programmer
  - Sarah Kell, editor
  - Caitlin Bird, research assistant
  - Ewa Czaykowska, compiler and chief editor

Map


British Columbia
Edmonton
Calgary
Vancouver
Washington
Montana
Portland
Oregon
Idaho
Wyoming
Nevada
Utah
Sacramento
Map data ©2013 Goo

The Nxaʔamxcín Language

- Southern Interior Salish; one of 7 Interior languages; one of 23 languages in the Salish family
- Spoken historically in north central Washington State
- Since 1870, confined to the Colville Reservation by Presidential Executive Order together with 8 other nomadic tribes
- Four varieties: Wenatchi, Moses-Columbia, Entiat and Chelan; roughly 3500 tribal members descended from Nxaʔamxcín-speaking families
- Fluent and active speakers in 2013: 2
- Nxaʔamxcín Language Program-since late 1990s

Initial purpose of project: 1991

- Based on Kinkade's fieldnotes and how he had transformed these notes into “data” on filecards
- To create a print dictionary
- Containing rich morphological information
- Organized according to root morphemes and lexicalized stems

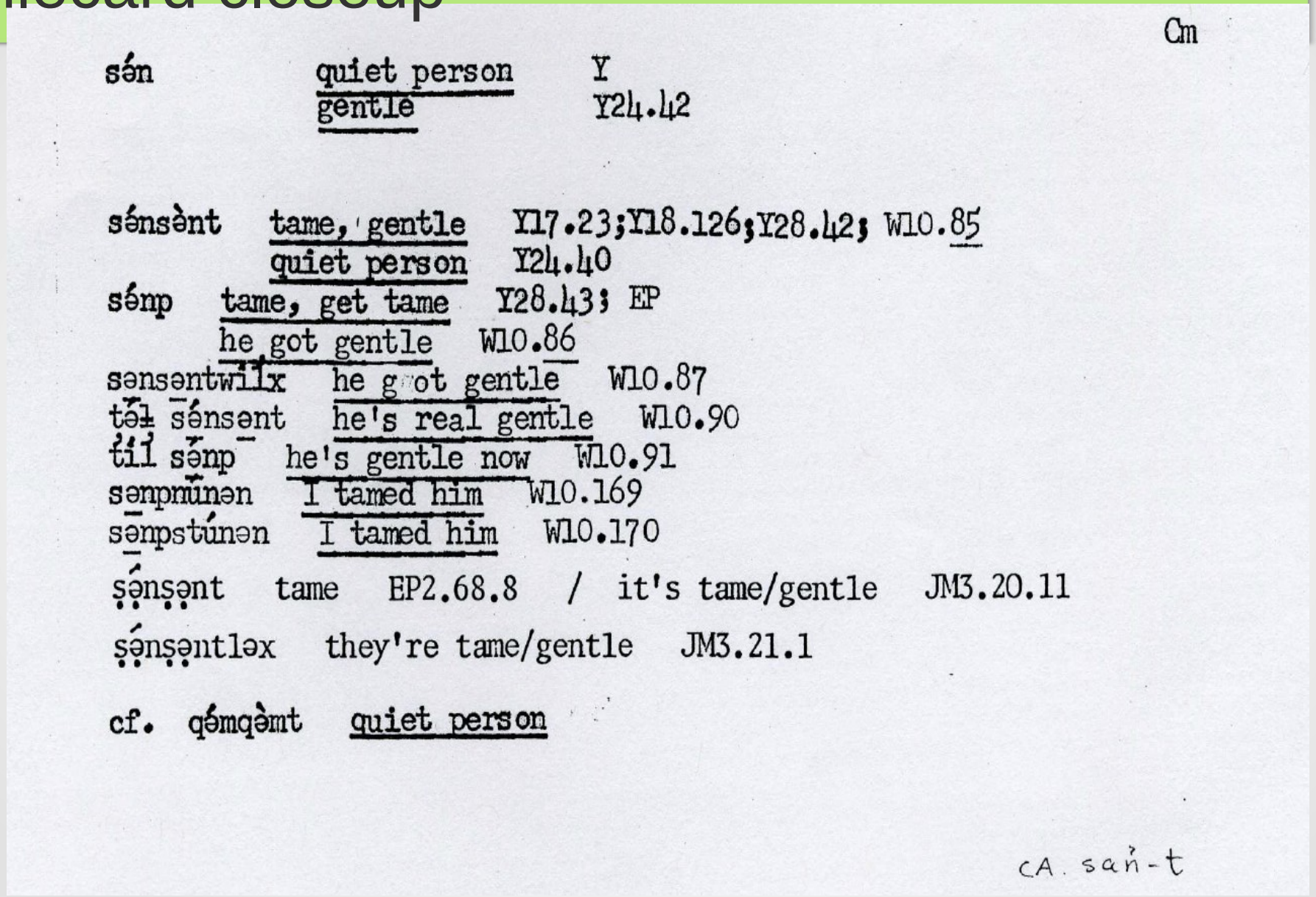
The initial target dictionary

- The entries were intended to look like the following from Kinkade's Upper Chehalis dictionary (1991)

628. **√k^{w}ənáˑ-**: hold, take, take hold, get, pick up, grab.
1. √k^{w}aná-mˈɬ, s√k^{w}aná-mˈal-n *detr.* **2.** FB s√k^{w}aná?-mɬ, -- *detr. pl.* **3a.** √k^{w}aná-n, s√k^{w}aná-t-n *tr.* ME --, kwâ-nât *tr.* **3b.** TC k^{w}anátn grasped, got hold. **3c.** TC ɬk^{w}anáčs *refl.* you'll take. **4.** FB √k^{w}anˈá-n or √k^{w}aná?-n, s√k^{w}an[a?]á-t-n or s√k^{w}an[?]á-t-n *tr. pl.* **5.** TCh FB ?ac√k^{w}án-ɬ he held. **6.** ME kwĭn-tûk hold. **7.** FB ?it √k^{w}aná?-n he held it for a while. **8.** √k^{w}ána-: catch, catch hold of, get, detain, FB capture. **8a.** √k^{w}ána-x^{w} or FB √k^{w}aná-x^{w}, s√k^{w}ána-y-n *tr.* **8b.** FB √k^{w}anˈá-x^{w}, s√k^{w}anˈá-y-n *tr. pl.* **9.** FB √k^{w}ána?-ɬ-čɬ she had,

The filecards

Filecard closeup



Cm

sén quiet person Y
gentle Y24.42

sénsènt tame, gentle Y17.23;Y18.126;Y28.42; W10.85
quiet person Y24.40
sénp tame, get tame Y28.43; EP
he got gentle W10.86
sənsəntwilx he got gentle W10.87
təɬ sénsənt he's real gentle W10.90
tíl sénp he's gentle now W10.91
sənpmúnən I tamed him W10.169
sənpstúnən I tamed him W10.170
ṣénṣənt tame EP2.68.8 / it's tame/gentle JM3.20.11
ṣénṣəntləx they're tame/gentle JM3.21.1

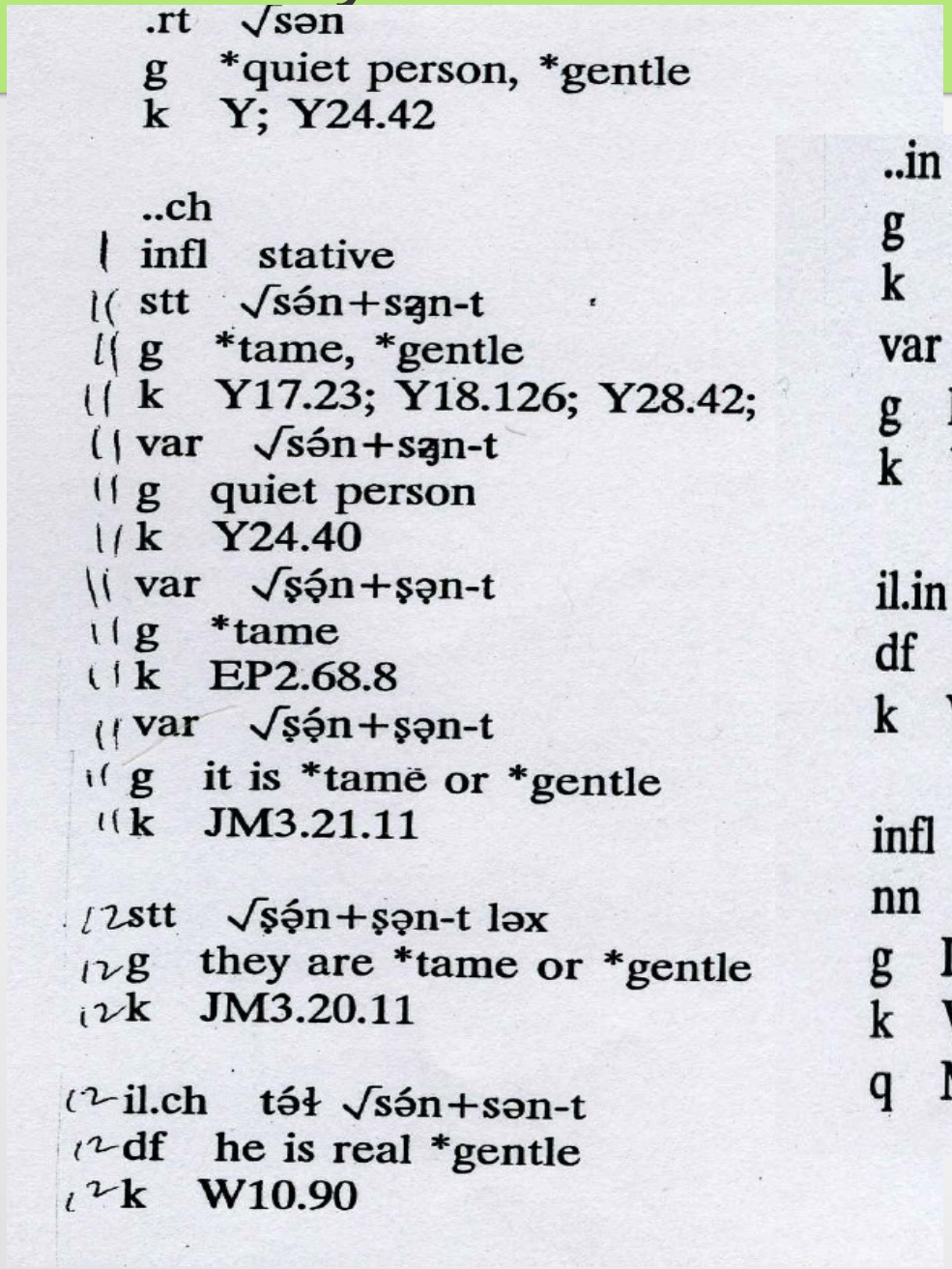
cf. qémqèmt quiet person

cA. saṅ-t

Filecards to Lexware

- In 1991 ….
- The “data” were input into a computer program called Lexware
- Lexware developed by Robert Hsu at U of Hawaii to create lexical resources
- Used band format, including bandnames, headwords, numbering and spaces to create a hierarchical system of embedding sub-entries within entries (Hsu 1985)

Lexware entry



.rt √sən
g *quiet person, *gentle
k Y; Y24.42

..ch
infl stative
stt √sə́n+sə̧n-t
g *tame, *gentle
k Y17.23; Y18.126; Y28.42;
var √sə́n+sə̧n-t
g quiet person
k Y24.40
var √ṣə́n+ṣən-t
g *tame
k EP2.68.8
var √ṣə́n+ṣən-t
g it is *tamë or *gentle
k JM3.21.11

stt √ṣə́n+ṣən-t ləx
g they are *tame or *gentle
k JM3.20.11

il.ch tə́ɬ √sə́n+sən-t
df he is real *gentle
k W10.90

..in √sə́n-p
g *tame, get tame
k Y28.43; EP
var √sə́n-p
g he got gentle
k W10.86

il.in t'ə́l' √sə́n-p
df he is *gentle now
k W10.91

infl success
nn √sən-p-nún-ən
g I tamed <*tame> him
k W10.169
q MDK underlined root schwa,

The data

```
.rt?    ÀEOTBELÀÀ1SOHÀkÀSOEOTÀ

linfl    transitive
l1tr    ÀEOTBELÀÀ1SOHÀÀUSSOHÀkÀSOEOTÀ©À9SOHÀn
l1g    *pull
q    should this be under a separate root?
```

The rescue

- Lexware binary data was dependent on:
  - WordPerfect for DOS
  - Customized character sets
  - Obsolete printer fonts
  - A Hercules graphics card
- AARGH! 3 months of misery. See Holmes & Newton (2008) for the whole horrible story.

Lessons learned

- Hardware goes away.
- Software goes away.
- Only your data remains.
- You'd better be able to read it.

Part II

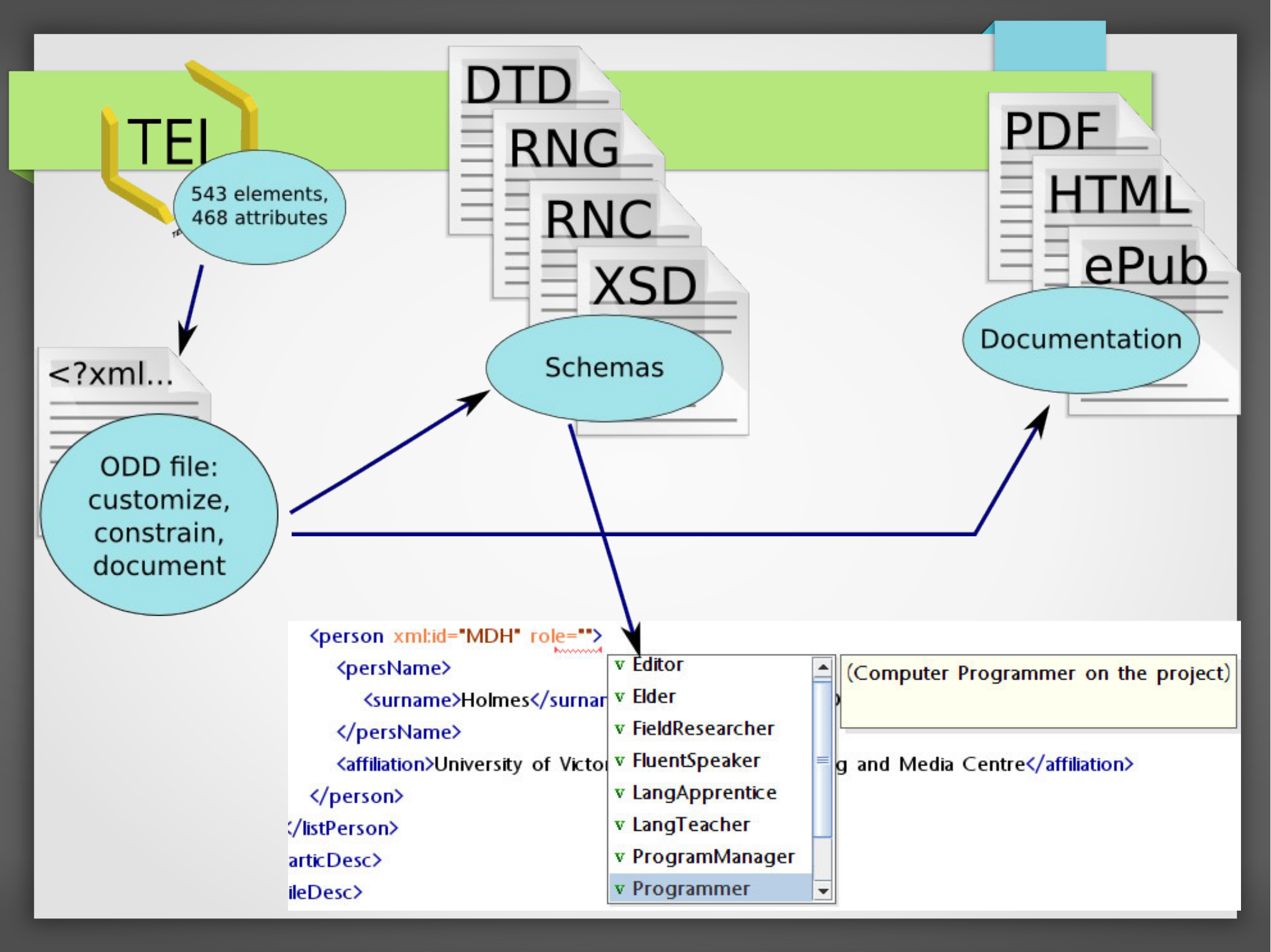
Using TEI for a Lexical Resource Project

Never again...

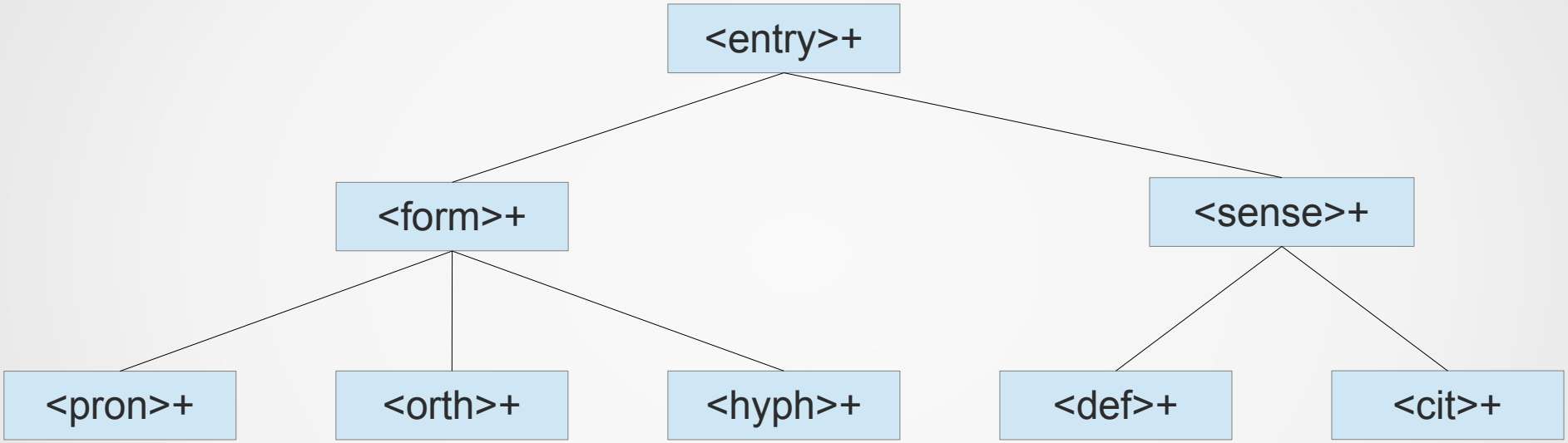
- From now on:
  - **Unicode** (stable, mature standard)
  - **XML** (stable, mature standard with conversion tools)
  - **TEI**

TEI: Text Encoding Initiative

- “...a consortium which collectively develops and maintains a standard for the representation of texts in digital form”
- Chief deliverables: schemas and a set of *Guidelines* which specify encoding methods for machine-readable texts
- Regular releases, thousands of projects, lots of support
- **Chapter 9: Dictionaries**:
  - Tags and attributes for encoding historical, print and born-digital lexica


TEI
543 elements,
468 attributes
DTD
RNG
RNC
XSD
Schemas
PDF
HTML
ePub
Documentation
<?xml...
ODD file:
customize,
constrain,
document
<person xml:id="MDH" role="">
<persName>
<surname>Holmes</surnar
</persName>
<affiliation>University of Victo
</person>
</listPerson>
articDesc>
ileDesc>
Editor
Elder
FieldResearcher
FluentSpeaker
LangApprentice
LangTeacher
ProgramManager
Programmer
(Computer Programmer on the project)
g and Media Centre</affiliation>

Structure of a TEI entry



Morpheme linking

- All morphemes have entries with feature structures:

```
<entry xml:id="ḥamʔ">
   <form>
     <pron>
       <seg type="p" subtype="i">ḥámʔ</seg>
       <bibl corresp="psn:ECH">ECH</bibl>
     </pron>
   </form>
   <sense>
     <def>
       <seg><gloss>forbid</gloss>, <gloss>stop</gloss></seg>
       <note type="editorial" resp="psn:ECH">Definition inferred based on complex forms containing this
root.</note>
     </def>
   </sense>
   <fs>
     <f name="baseType">
       <symbol value="root"/>
     </f>
   </fs>
   <note type="editorial" resp="psn:ECH">Root entry added based on attested complex forms.</note>
 </entry>
```

Morpheme linking

- All polymorphemic items are linked to the entries of their constituent morphemes:

```
<entry xml:id="ḥamʔcn">
   <form>
     <pron>
       <seg type="p" subtype="i">ḥámʔcn</seg>
       <bibl corresp="psn:ECH">ECH</bibl>
       <seg type="n">hámʔčən</seg>
       <bibl corresp="psn:JM">JM3.200.10</bibl>
     </pron>
     <hyph>
       √<m corresp="m:ḥamʔ">ḥámʔ</m>-<m corresp="m:stu">c</m>-<m corresp="m:Ø-OBJ m:n-
SUBJ">n</m>
     </hyph>
   </form>
  [...]
  </entry>
```

So we can do this:

/ḥámʔcn/ ECH [hámʔčən] JM3.200.10

Get related words

√ḥámʔ -c -n

COMPONENT MORPHEMES:

ḥámʔ | c | Ø-OBJ | n-SUBJ

/ḥámʔ/ (ROOT) ECH

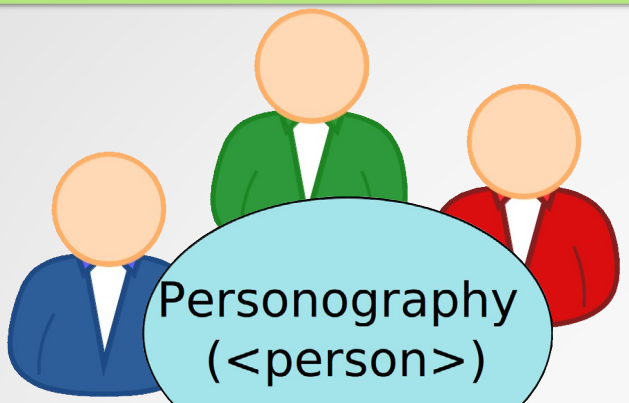
forbid, stop Note: Definition inferred based on complex forms containing this root.

Other entries containing this morpheme

ʔacḥámʔsn	stopped
ḥámʔcn	stopped
ḥámʔn	stop, forbid, forbade
ḥámʔn lx	stopped
hámʔntm lx	stopped

I stopped him from arguing/quarreling with someone JM3.200.10

TEI



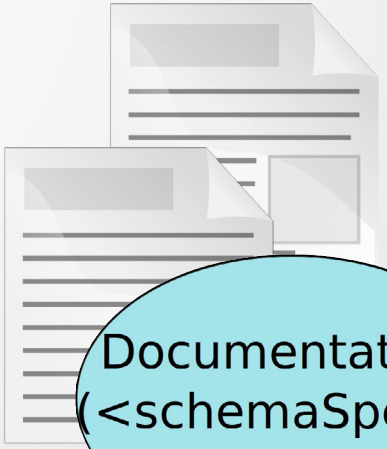
Personography
(<person>)

Placeography
(<place>,
<geo>)

Lexicon
entries
(<entry>)

Website pages
(, <p>...)

Audio/video
transcription
(<timeline>)



Documentation
(<schemaSpec>,
<valList>...)

Part III

Comparing Standards

Standardization and Interoperability

- Bird and Simons (2003) “Seven Kinds of Portability”

  “If digital language documentation and description should transcend time, they should …. be reusable... across different software and hardware platforms....and across different purposes.”

  - Interoperability is the ease of moving between systems/platforms
  - We use it here to refer to ease of moving between TEI and other dictionary-creating formats
  - Interoperability requires standardization of format and terminology

Format Interoperability

- XML-to-XML conversion is trivial.
- What schema should we aim at?
- **LMF version 16 DTD**?
  - 5 drawbacks at http://tla.mpi.nl/relish/lmf/.
- **LIFT**?
  - Supported by *WeSay*, *FLEx* and *Lexique Pro*, but last spec 2009, last code update over a year ago.
- **RELISH**?
  - Best candidate... Modular RNG schemas, and allows for TEI feature structures.

RELISH encoding

```
<LexicalEntry>
  <tei:f name="partOfSpeech"
    dcr:datcat="http://www.isocat.org/datcat/DC-1345"
    fVal="commonNoun"
    dcr:valueDatcat="http://www.isocat.org/datcat/DC-1256"/>
  <Lemma type="Form">
    <tei:f name="writtenForm"
      dcr:datcat="http://www.isocat.org/datcat/DC-1836"
      fVal="clergyman"/>
  </Lemma>
</LexicalEntry>
```

Terminological Interoperability

- RELISH depends on feature structures
- Feature structures depend on a shared ontology
- And that's a complication...

- One possible ontology is the GOLD ontology (General Ontology for Linguistic Description; http://linguistics-ontology.org), an attempt to create a stable, structured and documented standard
- GOLD is available through the ISOcat Data Category Registry which provides a formal system for associated terms from ontologies with feature structures in a specific encoding, using ISOcat URIs associated with specific concepts.

GOLD, ISOcat and Nxaʔamxcín

- Our feature structures are based on a morphological analysis of Nxaʔamxcín that itself is based on
  - 1) traditional Salishanist analysis as represented by the work of Kinkade;
  - 2) modified by subsequent analyses found in Willett (2003) and Czaykowska-Higgins (1998)
- This morphological analysis proposes a particular set of categories and relationships between the categories

(Mis)matches between ontologies

- We can match a small subset of our features straightforwardly onto the GOLD ontology
- There are also cases which have no match in GOLD but can be added
- But, there is a large set of cases (most) which map onto GOLD only loosely; crucially these mismatches arise from our specific morphological analysis
  - e.g., Distributive in Nxaʔamxcín involves plurality and events/actions repeated or distributed over time/space; the closest match is AugmentativeSize, but Distributive is not a Size category in Nxaʔamxcín

(Mis)matches between ontologies

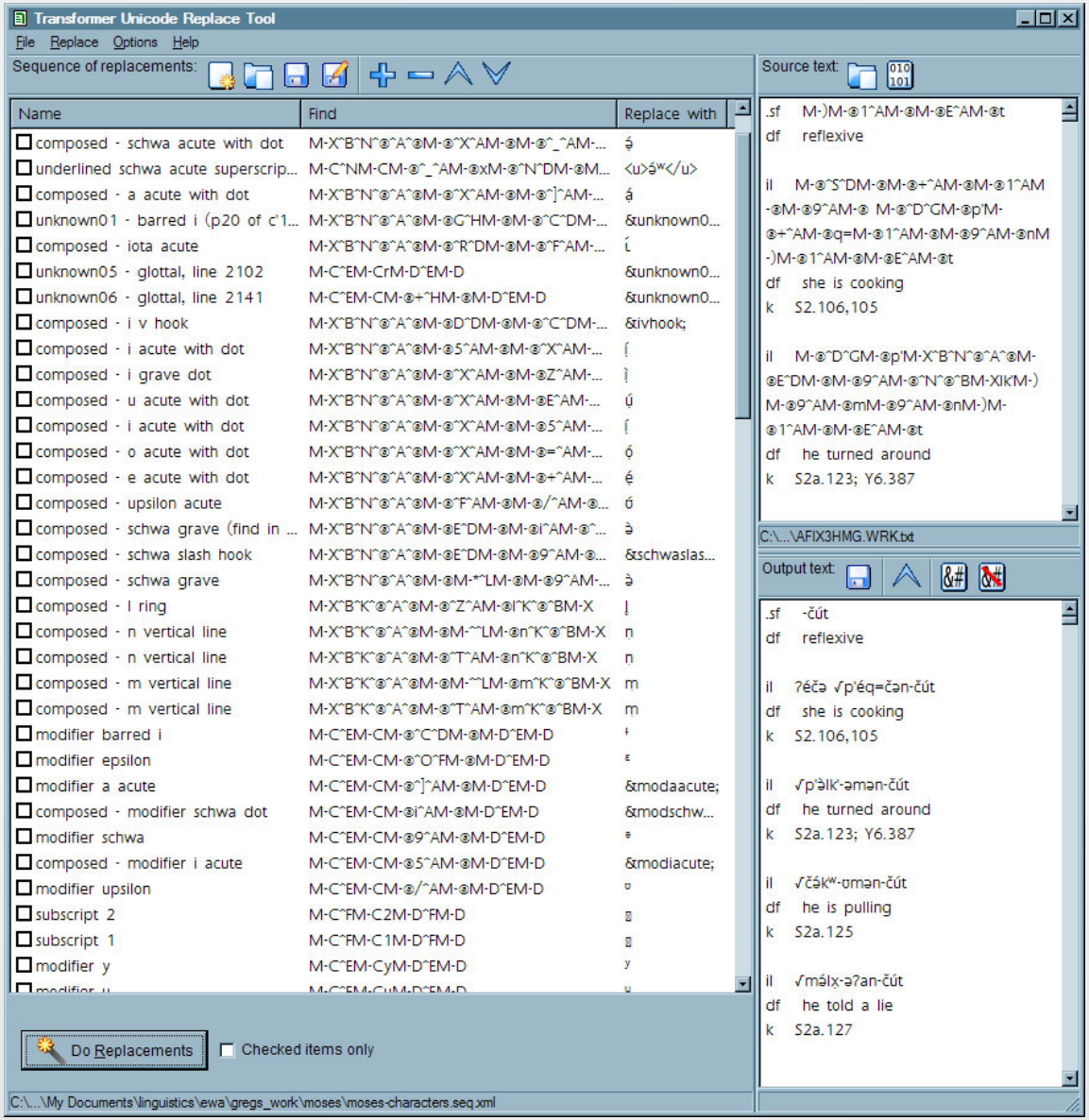
- Nxa'amxcín words have different suffixes that signal that stems are transitive or intransitive (e.g., applicatives); these are categorized in the Salishanist literature as 'Valence-changing' morphemes; GOLD categorizes many of these as Voice.
- The supracategories in GOLD do not align with the supracategories in our ontology.
  - e.g., GOLD's Voice properties (like applicative) are included in a supracategory 'Morphosyntactic'; its Transitivizer category is included in a supracategory 'Derivational'. But in the Nxaʔamxcín ontology all transitivizers are in one supracategory 'Morphosyntactic:Valence'
  - While it is not impossible to create a shared ontology, it is also not straightforward, and it is time-consuming

Conclusion

- Ultimately Interoperability is an important long term goal.
- Format conversion (XML → XML) is trivial, but...
- True interoperability depends on a shared ontology.
- The shared ontology depends on a shared grammatical analysis.
- This will take time to develop.
- In the meantime, in TEI, we have a stable, flexible schema system with solid documentation and a good toolset for creating the lexical resources we need now.

Lámlamt

To Unicode...


Transformer Unicode Replace Tool
File Replace Options Help
Sequence of replacements:
Name
Find
Replace with
composed - schwa acute with dot
M-X^B^N^@^A^@M-@^X^AM-@M-@^_^AM-...
ə́
underlined schwa acute superscrip...
M-C^NM-CM-@^_^AM-@xM-@^N^DM-@M...
<u>ə́ʷ</u>
composed - a acute with dot
M-X^B^N^@^A^@M-@^X^AM-@M-@^]^AM-...
á
unknown01 - barred i (p20 of c'1...
M-X^B^N^@^A^@M-@G^HM-@M-@^C^DM-...
&unknown0...
composed - iota acute
M-X^B^N^@^A^@M-@^R^DM-@M-@^F^AM-...
ί́
unknown05 - glottal, line 2102
M-C^EM-CrM-D^EM-D
&unknown0...
unknown06 - glottal, line 2141
M-C^EM-CM-@+^HM-@M-D^EM-D
&unknown0...
composed - i v hook
M-X^B^N^@^A^@M-@D^DM-@M-@^C^DM-...
&ivhook;
composed - i acute with dot
M-X^B^N^@^A^@M-@5^AM-@M-@^X^AM-...
í
composed - i grave dot
M-X^B^N^@^A^@M-@^X^AM-@M-@Z^AM-...
ì
composed - u acute with dot
M-X^B^N^@^A^@M-@^X^AM-@M-@E^AM-...
ú
composed - i acute with dot
M-X^B^N^@^A^@M-@^X^AM-@M-@5^AM-...
í
composed - o acute with dot
M-X^B^N^@^A^@M-@^X^AM-@M-@=^AM-...
ó
composed - e acute with dot
M-X^B^N^@^A^@M-@^X^AM-@M-@+^AM-...
é
composed - upsilon acute
M-X^B^N^@^A^@M-@^F^AM-@M-@/^AM-@...
ύ
composed - schwa grave (find in ...
M-X^B^N^@^A^@M-@E^DM-@M-@i^AM-@^...
ə̀
composed - schwa slash hook
M-X^B^N^@^A^@M-@E^DM-@M-@9^AM-@...
&schwaslas...
composed - schwa grave
M-X^B^N^@^A^@M-@M-*^LM-@M-@9^AM-...
ə̀
composed - l ring
M-X^B^K^@^A^@M-@^Z^AM-@l^K^@^BM-X
l̥
composed - n vertical line
M-X^B^K^@^A^@M-@M-^^LM-@n^K^@^BM-X
n
composed - n vertical line
M-X^B^K^@^A^@M-@^T^AM-@n^K^@^BM-X
n
composed - m vertical line
M-X^B^K^@^A^@M-@M-^^LM-@m^K^@^BM-X
m
composed - m vertical line
M-X^B^K^@^A^@M-@^T^AM-@m^K^@^BM-X
m
modifier barred i
M-C^EM-CM-@^C^DM-@M-D^EM-D
ɨ
modifier epsilon
M-C^EM-CM-@^O^FM-@M-D^EM-D
ɛ
modifier a acute
M-C^EM-CM-@^]^AM-@M-D^EM-D
&modaacute;
composed - modifier schwa dot
M-C^EM-CM-@i^AM-@M-D^EM-D
&modschw...
modifier schwa
M-C^EM-CM-@9^AM-@M-D^EM-D
ə
composed - modifier i acute
M-C^EM-CM-@5^AM-@M-D^EM-D
&modiacute;
modifier upsilon
M-C^EM-CM-@/^AM-@M-D^EM-D
ʊ
subscript 2
M-C^FM-C2M-D^FM-D
₂
subscript 1
M-C^FM-C1M-D^FM-D
₁
modifier y
M-C^EM-CyM-D^EM-D
ʸ
Do Replacements
Checked items only
C:\...\My Documents\linguistics\ewa\gregs_work\moses\moses-characters.seq.xml
Source text:
.sf M-)M-@1^AM-@M-@E^AM-@t
df reflexive
il M-@^S^DM-@M-@+^AM-@M-@1^AM-@M-@9^AM-@ M-@^D^GM-@p'M-@+^AM-@q=M-@1^AM-@M-@9^AM-@nM-)M-@1^AM-@M-@E^AM-@t
df she is cooking
k S2.106,105
il M-@^D^GM-@p'M-X^B^N^@^A^@M-@E^DM-@M-@9^AM-@^N^@^BM-XlK'M-)M-@9^AM-@mM-@9^AM-@nM-)M-@1^AM-@M-@E^AM-@t
df he turned around
k S2a.123; Y6.387
C:\...\AFIX3HMG.WRK.txt
Output text:
.sf -čút
df reflexive
il ʔéčə √p'éq=čən-čút
df she is cooking
k S2.106,105
il √p'əlK'-əmən-čút
df he turned around
k S2a.123; Y6.387
il √čə́kʷ-ʊmən-čút
df he is pulling
k S2a.125
il √mə́lχ-əʔan-čút
df he told a lie
k S2a.127

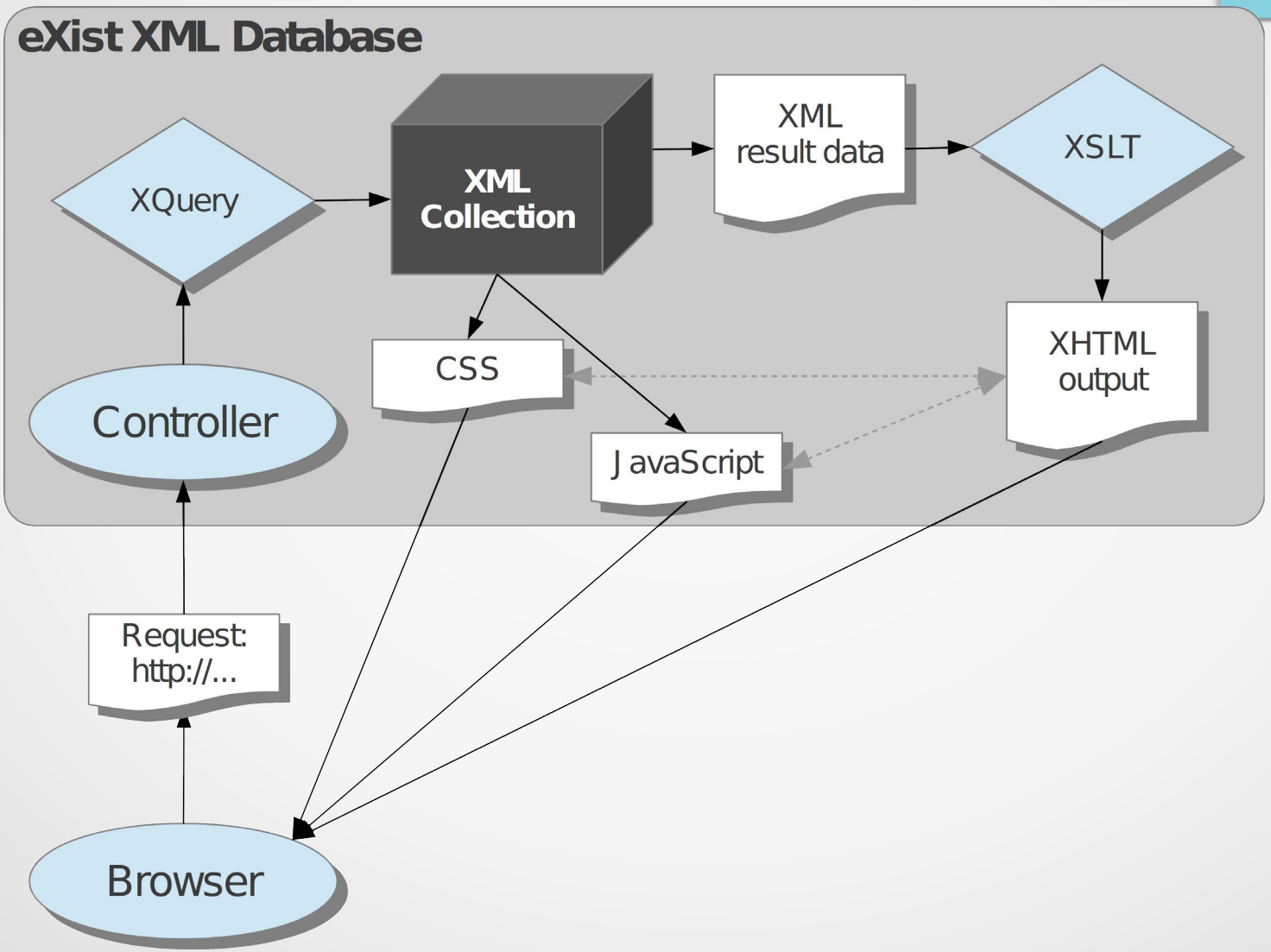
To XML...

```
<ENTRY level="001"  id="√cah">
 <rt>√cah</rt>
 <ENTRY level="002"  id="">
  <ls></ls>
  <infl mode="1">relational</infl>
  <mn mode="11">√cah=cí-mən</mn>
  <g mode="11">I *encourage|d someone</g>
  <k mode="11">W11.58</k>
  <var mode="11">√ca̱h=čé-mən</var>
  <g mode="11">I *remind|ed him</g>
  <k mode="11">JM3.119.5</k>
  <mn mode="12">√cah=cí-mən-c</mn>
  <g mode="12">he *encourage|d me</g>
  <k mode="12">W11.59</k>
  <mn.m mode="13">√ca̱h=čé-m-əm</mn.m>
  <g mode="13">I *remind|ed him</g>
  <k mode="13">Y41.125</k>
</ENTRY>
```

To TEI...

```
<entry xml:id="cahcimn">
  <form>
    <pron>
      <seg type="p" subtype="i">cahcímn</seg>
      <bibl corresp="psn:ECH">ECH</bibl>
      <seg type="n">cahcímən</seg>
      <bibl corresp="psn:W">W11.58</bibl>
    </pron>
    <hyph>√<m corresp="m:cah">cah</m>=<m corresp="m:cin">cí</m>-<m corresp="m:min m:t-TR m:Ø-OBJ m:n-SUBJ">mn</m>
    </hyph>
  </form>
  <sense>
    <def>
      <seg>I <gloss>encourage</gloss>d someone</seg>
      <bibl corresp="psn:W">W11.58</bibl>
    </def>
  </sense>
  <xr>See <ref target="m:c̓ahcimn">c̓ahcímn</ref>, and <ref target="m:nʔiʕWqn">nʔíʕʷqn</ref>.</xr>
</entry>
```

The Web application



The printouts